CLAIMS

What is claimed is:

1. A method for capitalizing text in a document, comprising:

processing a reference corpus to construct a plurality of dictionaries of capitalized terms, said plurality of dictionaries comprising a singleton dictionary and a phrase dictionary, where each record in the singleton dictionary comprises a word in lowercase, a range of phrase lengths m:n for capitalized phrases that the word begins, where m is a minimum phrase length and n is a maximum phrase length, and where each record in the phrase dictionary comprises a multi-word phrase in lowercase; and

adding proper capitalization to an input monocase document by capitalizing words found in mandatory capitalization positions; and by

looking up each word in the singleton dictionary and, if the word is found in the singleton dictionary, testing the corresponding phrase length range, where if the phrase length range indicates that the word does not start a multi-word phrase, capitalizing the word, while if the phrase length range indicates that the word does start a multi-word phrase, testing the word and an indicated plurality of next words as a candidate phrase to determine if the candidate phrase is found in the phrase dictionary and, if it is, capitalizing the words of the multi-word phrase.

2. A method as in claim 1, where if the candidate phrase is not found in the phrase dictionary, changing the number of words in the candidate phrase to form a revised candidate phrase and determining whether the revised candidate phrase is found in the phrase dictionary.

3. A method as in claim 1, wherein each record in said singleton dictionary further comprises an optional final form for the word if the word has unusual

capitalization, and where each record in said phrase dictionary comprises an optional final form for the phrase if the phrase has unusual capitalization.

4. A method as in claim 1, wherein when constructing the singleton dictionary all single word entities are added to the singleton dictionary with a phrase length range of 1:1, indicating that the word does not begin any phrase and should be capitalized by itself, and when constructing the phrase dictionary a multi-word phrase is added to the phrase dictionary, and the first word of the multi-word phrase is added to the singleton dictionary with a phrase length range of n:n, where n is the number of words in the phrase.

5. A method as in claim 4, wherein if the first word of the multi-word phrase already exists in the singleton dictionary, the phrase length range entry for the word is updated in the singleton dictionary so that the length of the current multi-word phrase is included in the phrase length range.

6. A method as in claim 1, wherein capitalizing words found in mandatory capitalization positions capitalizes words that begin sentences, words found in an abbreviations dictionary, and words found in a titles dictionary.

7. A method as in claim 6, wherein said abbreviations dictionary is constructed such that all words that end with a period at least X% of the time, and that precede a lower case word at least Y% of those times, are added to the abbreviations dictionary, and all words from the singleton dictionary that end with a period are added to the abbreviations dictionary.

8. A method as in claim 7, wherein X is about 75 and Y is about 50.

9. A method as in claim 1, wherein processing the reference corpus comprises:

counting a number of times each word in the reference corpus occurs lowercased (*l*), capitalized (*c*), all uppercase (*u*) and in a mandatory capitalization position (*m*); and

computing a capitalization probability *p* for each word in the reference corpus in accordance with:

$$p(C_i) = (c_i - m_i + u_i)/(l_i + c_i - m_i + u_i).$$

10. A method as in claim 9, and further comprising filtering named entities extracted from the reference corpus such that named entities that occur in fewer than Z documents are discarded, and all single-word named entities having a computed capitalization probability less than W are discarded; and storing surviving named entities into at least one of said plurality of dictionaries.

11. A method as in claim 10, wherein Z=3 and W=0.5.

12. A computer system for capitalizing text in a document, said computer system comprising a data processor that operates in accordance with program instructions recorded on an electronically readable program instruction storage media, said program instructions controlling said data processor for processing a reference corpus to construct a plurality of dictionaries of capitalized terms, said plurality of dictionaries comprising a singleton dictionary and a phrase dictionary, where each record in the singleton dictionary comprises a word in lowercase, a range of phrase lengths m:n for capitalized phrases that the word begins, where m is a minimum phrase length and n is a maximum phrase length, and where each record in the phrase dictionary comprises a multi-word phrase in lowercase; for adding proper capitalization to an input monocase document by capitalizing words found in mandatory capitalization positions; and for looking up each word in the singleton dictionary and, if the word is found in the singleton dictionary, for testing the corresponding phrase length range, where if the phrase length range indicates that the word does not start a multi-word

phrase, capitalizing the word, while if the phrase length range indicates that the word does start a multi-word phrase, for testing the word and an indicated plurality of next words as a candidate phrase to determine if the candidate phrase is found in the phrase dictionary and, if it is, capitalizing the words of the multi-word phrase.

13. A computer system as in claim 12, where said program instructions further control said data processor so that, if the candidate phrase is not found in the phrase dictionary, for changing the number of words in the candidate phrase to form a revised candidate phrase and for determining whether the revised candidate phrase is found in the phrase dictionary.

14. A computer system as in claim 12, wherein each record in said singleton dictionary further comprises an optional final form for the word if the word has unusual capitalization, and where each record in said phrase dictionary comprises an optional final form for the phrase if the phrase has unusual capitalization.

15. A computer system as in claim 12, where said program instructions further control said data processor, when constructing the singleton dictionary, such that all single word entities are added to the singleton dictionary with a phrase length range of 1:1, indicating that the word does not begin any phrase and should be capitalized by itself, and when constructing the phrase dictionary a multi-word phrase is added to the phrase dictionary, and the first word of the multi-word phrase is added to the singleton dictionary with a phrase length range of n:n, where n is the number of words in the phrase.

16. A computer system as in claim 15, where said program instructions further control said data processor, if the first word of the multi-word phrase already exists in the singleton dictionary, such that the phrase length range entry for the word is updated in the singleton dictionary so that the length of the current multi-word phrase is included in the phrase length range.

17. A computer system as in claim 12, where said program instructions further control said data processor, when capitalizing words found in mandatory capitalization positions, to capitalize words that begin sentences, words found in an abbreviations dictionary, and words found in a titles dictionary.

18. A computer system as in claim 12, where said program instructions further control said data processor, when processing the reference corpus, for counting a number of times each word in the reference corpus occurs lowercased (*l*), capitalized (*c*), all uppercase (*u*) and in a mandatory capitalization position (*m*); and for computing a capitalization probability *p* for each word in the reference corpus in accordance with:

$$p(\mathrm{C}_i) = (c_i - m_i + u_i)/(l_i + c_i - m_i + u_i).$$

19. A computer system as in claim 18, where said program instructions further control said data processor for filtering named entities extracted from the reference corpus such that named entities that occur in fewer than Z documents are discarded, and all single-word named entities having a computed capitalization probability less than W are discarded; and for storing surviving named entities into at least one of said plurality of dictionaries.

20. A computer system as in claim 19, wherein Z=3 and W=0.5.